CTrGAN: Cycle Transformers GAN for Gait Transfer

Shahar Mahpod   Noam Gaash   Hay Hoffman   Gil Ben-Artzi

Ariel University *

Israel

http://gil-ba.com

Abstract

We introduce a novel approach for gait transfer from unconstrained videos in-the-wild. In contrast to motion transfer, the objective here is not to imitate the source's motions by the target, but rather to replace the walking source with the target, while transferring the target's typical gait. Our approach can be trained only once with multiple sources and is able to transfer the gait of the target from unseen sources, eliminating the need for retraining for each new source independently. Furthermore, we propose a novel metrics for gait transfer based on gait recognition models that enable to quantify the quality of the transferred gait, and show that existing techniques yield a discrepancy that can be easily detected.

We introduce Cycle Transformers GAN (CTrGAN), that consist of a decoder and encoder, both Transformers, where the attention is on the temporal domain between complete images rather than the spatial domain between patches. Using a widely-used gait recognition dataset, we demonstrate that our approach is capable of producing over an order of magnitude more realistic personalized gaits than existing methods, even when used with sources that were not available during training. As part of our solution, we present a detector that determines whether a video is real or generated by our model.

1. Introduction

The goal of this paper is to provide both a novel approach and a detection mechanism for gait transfer from videos in-the-wild. The objective is to replace a walking person (source) in a video sequence with photorealistic images of a different walking person (target), such that the resulting gait is identifiable as the target's while still mimicking the source's basic motion.

Our approach learns directly from an unpaired collection of unconstrained videos in-the-wild containing walking people. We avoid the need for paired data and the need to disentangle the walking patterns into different representations and learn directly from the 2D frames. We train our model to translate multiple sources to a single target, so at inference time it can generalize to unseen sources without the need for retraining.

Motion transfer methods aim to synthesize a video in which one individual is acting in accordance with that of a different individual in a given real video. A growing body of research has been conducted on this topic, which has led to the development of advanced detection [33, 27, 10] and enhanced motion transfer techniques [44, 10]. As a result of their mutually reinforcing relationship, motion transfer technology can produce convincingly realistic images and videos through deep learning-based manipulations. Using the whole-body motion transfer approach that directly works on unconstrained videos [10] for gait transfer has the following key limitation: it attempts to replicate the precise movements of the source; rather, the goal of gait transfer is to translate the typical motions and appearances of the source into those of the target, adjusting for varied angles, paces, and shapes. To address this limitation we introduce CTrGAN for gait transfer. It transfers a series of poses from the source to the target while maintaining the natural movements of the target. Sources may vary in viewpoint, shape, and pace. It is based on Transformers [41], which have proven to be successful in translation tasks. Similar to NLP's models [47, 7], each Transformer consists of an encoder and a decoder. As a result, we can successfully translate between the sequences of poses of the sources and the targets. Our Transformer model performs self- and cross-attention in time rather than in image space, capturing the dynamic of the object. In order to generate unseen natural poses of the target, our model is trained in an unsupervised manner on unpaired data. This is in contrast to prior whole-body motion transfer approaches that required paired data (e.g., [43]). Figure 1 shows our method.

The quality of whole-body motion transfer is often evaluated in a supervised manner, based on the ability to approximate a pose and appearance of the target unseen during
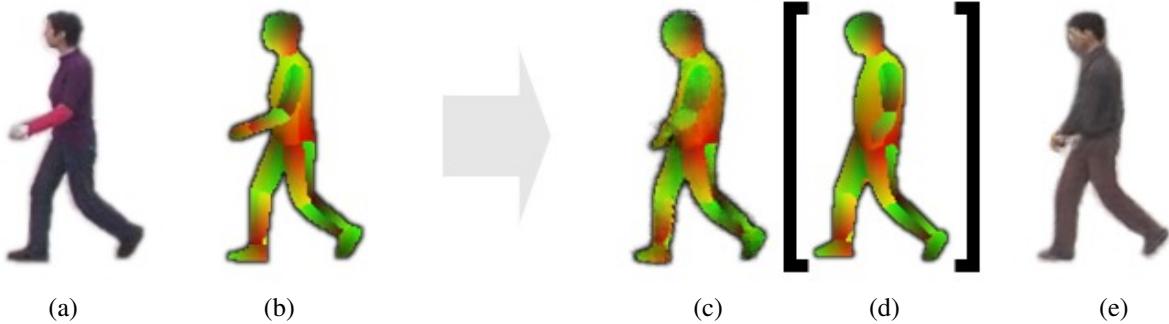
Figure 1: CTrGAN transfers the poses of the source to the target, while maintaining the natural gait of the target. From left to right: (a) The source's image is converted to (b) DensePose's [20] IUV format. (c) Our model translates the IUV of the source to the corresponding most natural IUV pose of the target by synthesizing a novel pose. (d) The generated pose is very similar (but not identical) to an exiting *real* pose in the dataset. (e) The generated pose is rendered to a corresponding RGB image of the target.

training but available in the test set. In our case, in addition to the appearance, the objective is to measure the translation of the gait pattern (i.e. the dynamic) from the source to a typical gait pattern of the target. In many cases, there are no one-to-one correspondences between each frame in the newly generated dynamic of the target and the existing ones already included in the dataset, and therefore it can not be evaluated directly in a supervised manner. We propose employing state-of-the-art gait recognition algorithms [38] to evaluate the quality of the gait transfer. The quality is determined by the accuracy with which newly generated poses are recognized as the target's gait. In order to provide accurate measurements in all scenarios, we use several different algorithms [11, 16, 28]. As a way of assessing the appearances of unseen poses, Chamfer's distance [6] is used.

Our model includes two networks: (a) CTrGAN, which translates the poses of the sources to the poses of the target, and (b) pose-to-appearance, which renders the appearance of each pose. For the latter network, we deploy an independent state-of-the-art existing approach.

This paper contributes by (a) introducing an approach for gait transfer from unconstrained videos in-the-wild, as well as evaluation metrics; (b) presenting Cycle Transformers GAN with temporal attention that can generate realistic gait patterns of the targets and a corresponding detector; and (c) demonstrating the effectiveness of our approach based on a standard gait recognition dataset, showing that it can generalize to unknown input sources, yielding the desired gait in an order of magnitude more cases than previous methods.

2. Related work

Pose-to-Pose/Appearance A variety of methods have been introduced for the generation of video sequences of the target based on semantic input, including facial mo-

tion transfer [35, 26, 4, 40] and whole-body motion transfer [10, 12, 43, 44, 32, 48]. These methods are based on the ability to accurately estimate the pose [20, 8] and also on image-to-image translation models [45, 51]. They are either explicitly trained for each source [10] or can be trained only once [43] as our approach. In contrast to previous works, our goal is to generate the personalized gait pattern of the target to best match the gait of the source rather than to accurately imitate the original motion of the source. Building on the recent advances, we employ [10, 44] as our pose-to-appearance network where the input is the generated poses of the target and not the source's poses. Our experiments demonstrate the benefits of CTrGAN over the direct use of [10, 44] for gait transfer. In the context of computer animation, [1] introduced motion style transfer. However, they require a separation between the walking style and its content and 3D joint positions whereas we learn directly from the images containing the walking persons without disentangled representations.

Gait Recognition. In recent years, various works have been proposed that use neural network models to identify people based on their gait [11, 16, 39, 13]. GaitSet [11] considers the gait as a set consisting of independent frames and recognizes it based on a sequence of silhouette images. GaitGL [28] relay on both global visual information and local region details and introduced attention between adjacent frames. GaitPart [16] uses a novel part-based model to characterize the gait. We use GaitSet, GaitGL, and Gait-Part models to assess the quality of the generated video sequence. When the source in the video sequence is replaced with the target, the identified gait should be replaced as well. We show that for previous approaches, a gait is still readily associated with its source while using our approach it is considered to belong to the target.

Visual Transformers. Transformers [41] are proven architectures in the field of Natural Language Processing [7], and several works have been done in recent years to adapt them to computer vision [30, 14, 23, 29]. The Transformer model is shown in [41] consists of two main components: an encoder and a decoder, which jointly process the input sequence, based on the self-attention mechanism. Early works [30] adapted Transformers to the image domain. Even though this work demonstrated its ability only on very small images, it paved the way for broader works that addressed common challenges such as object detection [9] and classification [14]. Recently, several works [23, 49, 24] have been presented which show that Transformers can also be incorporated into the GAN architecture for image generation tasks. Unlike previous approaches, our method transfers motion between domains cyclically using unpaired data [51, 5] and is based on attention in the temporal domain.

3. Method

CTrGAN differs in the following ways from CycleGAN and standard Transformer-based architecture. First, unlike CycleGAN, it cycles between domains by using a series of images rather than individual images. Second, unlike CycleGAN and Transformers, the attention is on the temporal domain between consecutive images and not between patches of the same image. This allows us to incorporate the target's gait pattern into the source's gait pattern transfer process. Third, we do not use positional encoding due to the approximate cyclic pattern of gait. Figure 2 depicts a schematic illustration of our Natural Gait Retargeting approach.

3.1. CTrGAN Architecture

The Cycle Transormer GAN (CTrGAN) consists of three main ingredients: features extractors, Transformers, and a cyclic process. We denote $\mathcal{I} = \{\mathbf{I}_j\}_{j=1}^M$ as a collection of RGBA images and $\mathcal{P} = \{\mathbf{P}_j\}_{j=1}^N$ as a collection of IUV images [20]. $\mathbf{I}_j^{s_i}, \mathbf{P}_j^{s_i}$ and $\mathbf{I}_j^t, \mathbf{P}_j^t$ denote the corresponding j^{th} image of the source and the target from the corresponding collection, respectively. In the following, we describe the details with respect to the target. Details regarding the source are derived in a similar manner. Below, the values in parenthesis represent those that we use in our implementation.

3.1.1 Transformers.

The Transformers follow the same architecture as presented in [41]. Originally, Transformers were designed to handle sequences and consisted of two components, an encoder, and a decoder. The encoder is designed to handle information that remains constant throughout the series, while the

decoder is designed to handle the continuous flow of information. The encoder and decoder consist of several chained attention blocks, and each receives three types of data as input: Keys, Values, and Queries (hereafter K, V, and Q).

3.1.2 Keys

The Keys of the target are selected from the images' collection whose feature vectors are the closest to the cluster centers. Let U_j denote the feature vector of image j:

$$\{\mathbf{U}_j\} = PCA(VGG_{16}(\{\mathbf{P}_j^t\}_{j=1..N}), d), \quad (1)$$

where VGG_{16} [37] are the features of the last layer (classifier) of the pre-trained VGG [1] [37, 31] and d is the PCA dimension (d = 100).

C_k denotes the center of the clusters as obtained by the K-means clustering:

$$\{\mathbf{C}_k\} = K_{means}(\{\mathbf{U}_j\}, m), \quad (2)$$

where m is the number of centroids (m = 18). Finally, the Keys $\{\mathbf{K}_k^t\}$ are defined as:

$$\mathbf{K}_k^t = \{\mathbf{P}_{p_k}^t\} \quad , \quad p_k = \{\underset{j}{\operatorname{argmin}} \|\mathbf{U}_j - \mathbf{C}_k\|_2\}. \quad (3)$$

Samples of the Keys that were used for one of the subjects can be seen in Figure 4a.

3.1.3 Queries

Given an input sequence $\mathbf{P}^t$, the queries are a sliding window of l_w consecutive frames ($l_w = 3$). We begin with a given frame and advance one sample at a time. During training, we choose a subsequence of length L at random and begin with its first frame. We process the entire series from beginning to end during inference.

3.1.4 Features Encoders and Decoders

Given an IUVA image $\mathbf{P}_j^t \in \mathbb{R}^{4 \times H \times W}$, where IUVA is an IUV image with an additional alpha layer, we generates a feature tensor $\mathbf{U}_j^t \in \mathbb{R}^{256 \times \frac{H}{64} \times \frac{W}{64}}$. The feature encoder $\mathcal{F}_T$ is a 5-layers CNN followed by 4 strides *max-pooling* (see Supplementary for more details).

The CTrGAN model includes two pairs of encoders ($\mathcal{F}_T$ and $\mathcal{F}_S$), where each of the pairs shares weights. All four feature encoders (shown in Figure 3) have the same structure. The $\mathcal{H}_T$ and $\mathcal{H}_S$ decoders are identical to the $\mathcal{F}_T$ encoder, except that they operate in the other direction. The discriminators $D_{t \to s}$ and $\mathcal{D}_{t \to s}$ (Figure 3) are 5-layers CNN (see Supplementary for more details).

---

[1] Trained on ImageNet-1K
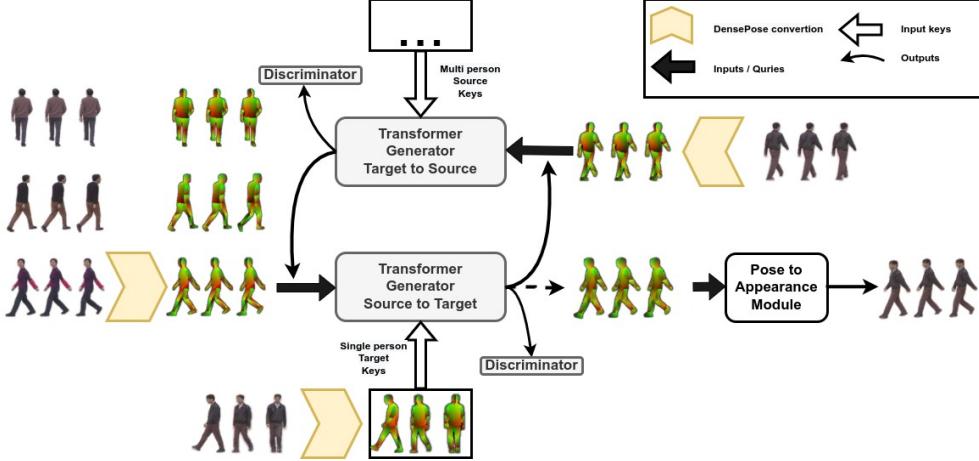
Figure 2: The generators of CTrGAN are based on transformers. The inputs to each generator are IUVA gait images from the training set and Keys. The outputs are natural gait poses. See the text for further details.
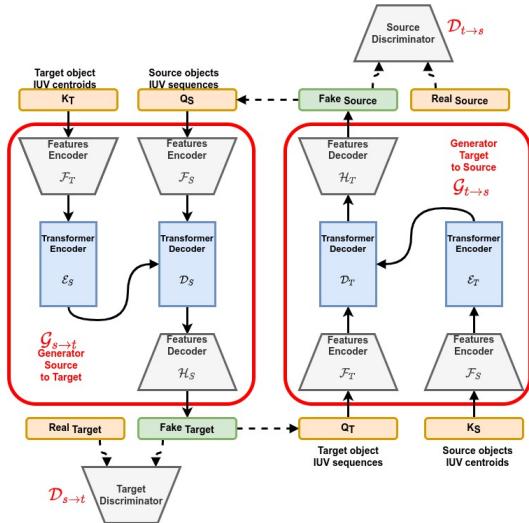

Figure 3: CTrGAN consists of two branches that are connected cyclically, feature encoders and decoders and Transformers which perform self- and cross-attention between the features.

3.1.5 Cycle Transformer GAN

Figure 3 shows the architecture of CTrGAN . We denote the source images collections as $\Phi^{s_i} = \{\mathbf{I}^{s_i}, \mathbf{P}^{s_i}\}_{i=1..N}$ and the target image collection as $\Psi^t = \{\mathbf{I}^t, \mathbf{P}^t\}$. We define two networks $\mathcal{G}_{s \to t}$ and $\mathcal{G}_{t \to s}$. The first network adapts pose images from a variety of sources to pose images of the target, while the second network does the opposite. We denote $\widetilde{\mathbf{P}}^t$ and $\widetilde{\mathbf{P}}^s$ as the outputs of the networks:

$$\left\{\widetilde{\mathbf{P}}^t\right\} = \mathcal{G}_{s \to t}\left(\{\mathbf{K}^t\}, \{\mathbf{P}^{s_i}\}\right), \tag{4}$$

$$\left\{\widetilde{\mathbf{P}}^{s_i}\right\} = \mathcal{G}_{t \to s}\left(\{\mathbf{K}^{s_i}\}, \{\mathbf{P}^t\}\right). \tag{5}$$

For brevity, images' indices have been omitted. Using the pose images of the source's gait pattern, our method attempts to generate pose images of the target while preserving its gait pattern. This domain adaptation is accomplished by composing both $\mathcal{G}_{s \to t}$ and $\mathcal{G}_{t \to s}$ in a cyclic manner as shown in the following:

$$\widetilde{\mathbf{P}}^{s_i} = \mathcal{G}_{t \to s}\left(\mathbf{K}^{s_i}, \mathcal{G}_{s \to t}\left(\mathbf{K}^t, \mathbf{P}^{s_i}\right)\right), \tag{6}$$

$$\widetilde{\mathbf{P}}^t = \mathcal{G}_{s \to t}\left(\mathbf{K}^t, \mathcal{G}_{t \to s}\left(\mathbf{K}^{s_i}, \mathbf{P}^t\right)\right). \tag{7}$$

The output pose image $\widetilde{\mathbf{P}}^t$ (*i.e.* $\mathcal{G}_{s \to t}\left(\mathbf{K}^t, \mathbf{P}^{s_i}\right)$)are used to generate the requested appearance by pose-to-appearance network $\mathcal{G}_M$:

$$\widetilde{\mathbf{I}}_i^t = \mathcal{G}_M\left(\mathbf{P}_i^t\right). \tag{8}$$

3.1.6 Self and Cross-Attention

The attention layer at the attention block is one of the core components of the Transformer. A detailed explanation of the self- and cross-attention and visualizations can be found in the Supplementary Materials. Figure 4b visualizes the attention mechanism.

3.1.7 Gait Cycle

We do not use positional encoding due to the cyclic pattern of gait [36]. The gait cycle can be defined as the time

(a) Keys
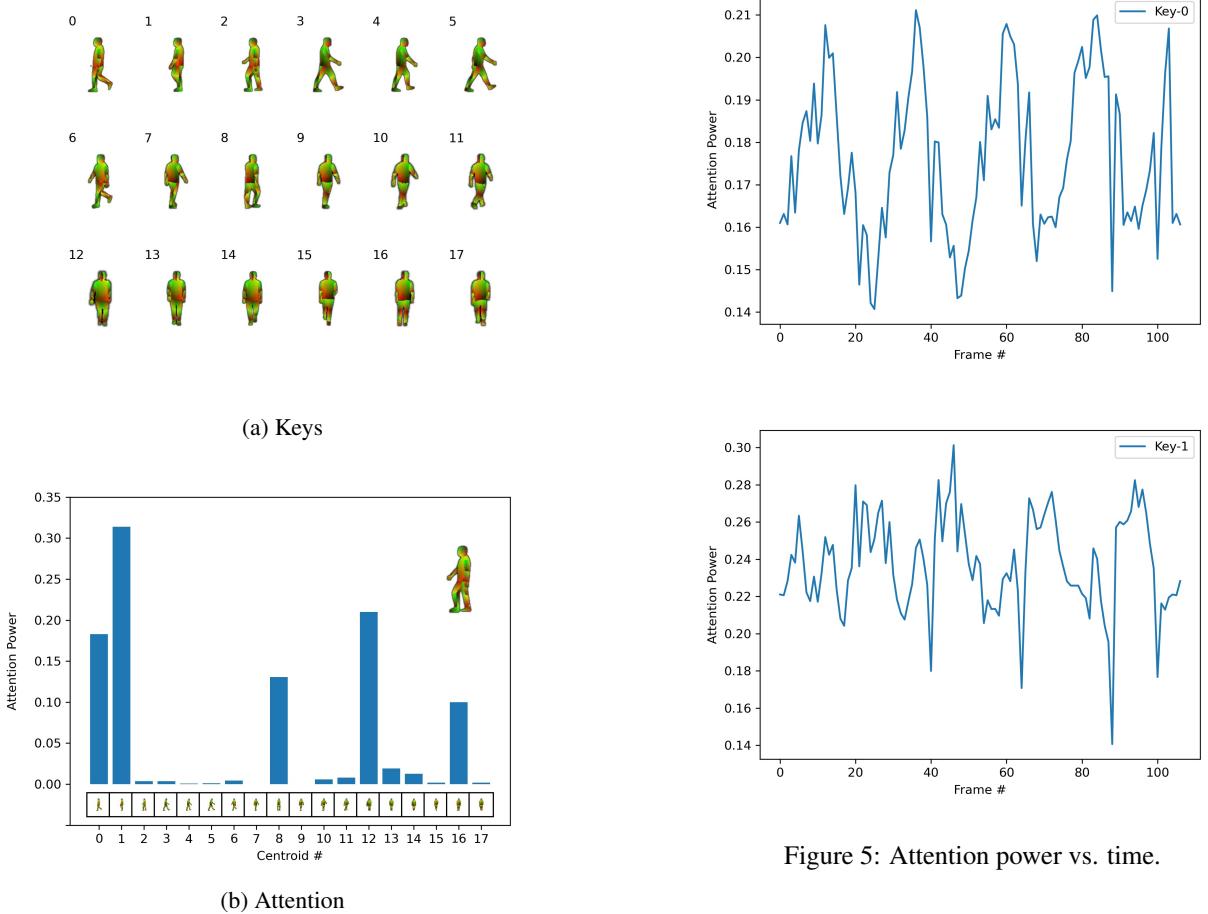
(b) Attention

Figure 4: (a) Samples of the Keys (centroids) $\mathbf{K}_k^t$ that were used. (b) A visual demonstration of our attention mechanism.

interval between two successive occurrences of one of the repetitive phases of locomotion [2]. Here, we demonstrate the periodicity of the movement as expressed by the cross-attention patterns along time for the Keys. Figure 5 shows the cross-attention of Key 0 (top) and Key 1 (bottom) over time. As can be seen, the gait cycle is evident.

3.2. Optimization and Loss Functions

We use the following loss functions in our training:

$$\ell_{cycle} = \lambda_{idt}\ell_{idt} + \lambda_{adv}\ell_{adv} + \lambda_{cyc}\ell_{cyc} + \lambda_{per}\ell_{per}, \quad (9)$$

where λ_{idt}, λ_{adv}, λ_{cyc} and λ_{per} are the weights of the losses. The identity loss function ℓ_{idt} is used to ensure that the cyclic mapping preserves the mapping from a pose image to itself. We use $\mathcal{L}_1$ loss function.

$$\ell_{idt} = \mathcal{L}_1\left(\mathcal{G}_{s\to t}\left(\mathbf{P}^t\right), \mathbf{P}^t\right) + \mathcal{L}_1\left(\mathcal{G}_{t\to s}\left(\mathbf{P}^s\right), \mathbf{P}^s\right). \quad (10)$$
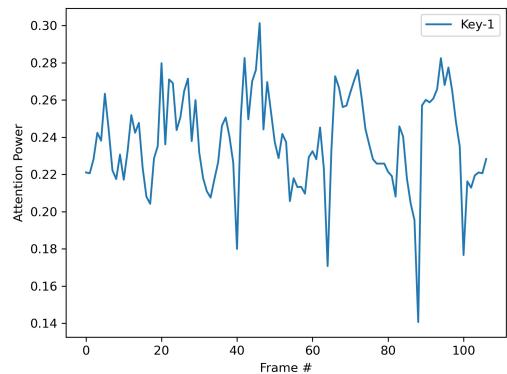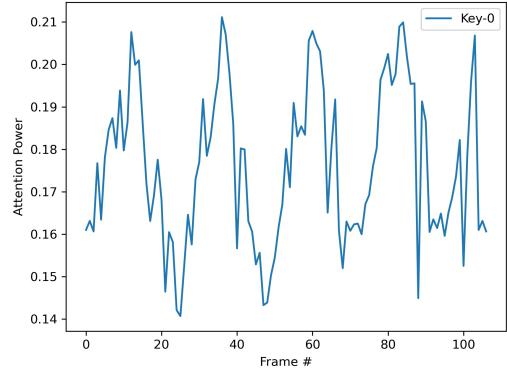

Figure 5: Attention power vs. time.

Following CycleGAN [51] and ReCycleGAN [5] we use the same adversarial loss ℓ_{adv}, in which two discriminator networks $\mathcal{D}_{s\to t}$ and $\mathcal{D}_{s\to t}$ are learned as a part of the training process.

In the same manner as GAN [19] architectures, we use the generator and discriminator $\mathcal{G}_{s\to t}$, $\mathcal{G}_{t\to s}$ $\mathcal{D}_{s\to t}$ and $\mathcal{D}_{t\to s}$. In our training process, we use the $\mathcal{L}_2$ as the objective loss function of the adversarial loss ℓ_{adv}.

$$\begin{aligned}\ell_{adv} = &\mathcal{L}_2\left(\mathcal{D}_{s\to t}\left(\mathcal{G}_{s\to t}\left(\mathbf{P}^s\right)\right), \mathbb{0}\right) + \mathcal{L}_2\left(\mathcal{D}_{s\to t}\left(\mathbf{P}^t\right), \mathbb{1}\right)\\ &+\mathcal{L}_2\left(\mathcal{D}_{t\to s}\left(\mathcal{G}_{t\to s}\left(\mathbf{P}^t\right)\right), \mathbb{0}\right) + \mathcal{L}_2\left(\mathcal{D}_{t\to s}\left(\mathbf{P}^s\right), \mathbb{1}\right)'\end{aligned}$$

(11)

where $\mathbb{0}$ and $\mathbb{1}$ are matrices of zeros and ones with the same dimensions as $\mathcal{D}_{s\to t}\left(\mathcal{G}_{s\to t}\left(\mathbf{P}^s\right)\right)$ and $\mathcal{D}_{t\to s}\left(\mathcal{G}_{t\to s}\left(\mathbf{P}^t\right)\right)$.

The cycle loss function ℓ_{cyc} which is the main core of the cycle GAN process is defined as:

$$\begin{aligned}\ell_{cyc} = &\mathcal{L}_2\left(\mathcal{G}_{t\to s}\left(\mathcal{G}_{s\to t}\left(\mathbf{P}^s\right)\right), \mathbf{P}^s\right)\\ &+\mathcal{L}_2\left(\mathcal{G}_{s\to t}\left(\mathcal{G}_{t\to s}\left(\mathbf{P}^t\right)\right), \mathbf{P}^t\right).\end{aligned}$$

(12)

The perceptual loss ℓ_{per} measures the difference between feature vectors of a real image and a generated one:
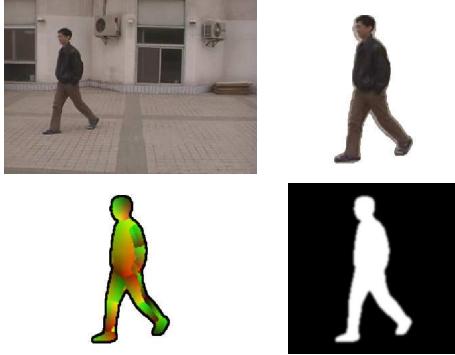
Figure 6: Pre-processing procedure - From left to right: 1. Original image. 2. Cropped and centered image. 3. Pose image (IUV - DensePose format). 4. Masked image, created from the I part of the IUV image

$$\ell_{per} = \mathcal{L}_1\left(VGG_{16}\left(\mathcal{G}_{t\rightarrow s}\left(\mathcal{G}_{s\rightarrow t}\left(\mathbf{P}^s\right)\right)\right), VGG_{16}\left(\mathbf{P}^s\right)\right)$$
$$+\mathcal{L}_1\left(VGG_{16}\left(\mathcal{G}_{s\rightarrow t}\left(\mathcal{G}_{t\rightarrow s}\left(\mathbf{P}^t\right)\right)\right), VGG_{16}\left(\mathbf{P}^t\right)\right). \tag{13}$$

We use a pre-trained VGG model [37, 31] to extract the feature vectors. VGG model is pre-trained on three layers of RGB, whereas we use four layers (IUVA), we therefore measured the perceptual loss of the IUV and the alpha channels separately.

4. Experiments

4.1. Dataset

We use CASIA-A [42], which is a widely-used gait recognition dataset [25, 15, 3, 18]. It includes twenty subjects. Each subject has twelve image sequences that were captured from three different viewpoints, resulting in four instances for each of the viewpoints (denoted by 001, 002, 003 and 004). Overall, there are a total of 240 video sequences with a resolution of 352x240 and 19,139 images.

We remove the background from the original CASIA-A images by using DensePose [20]. The input to our model is four-channel images (RGBA). In each frame, we extract the binary mask for the subject and attach it to the RGB image as an alpha channel. The images are cropped and centered around the object to create 256x256 canvases. See Figure 6 for examples.

4.2. Implementation Details

The networks are implemented using Pytorch [31] and were trained on a single NVidia 2080Ti GPU. We train the model with Adam optimizer with $\beta 1 = 0.5$ $\beta 2 = 0.999$ over 20 epochs. The initial learning rate is set to $2e-4$, for 5 epochs followed by a linear decay to zero over 15

more epochs. The same configurations and parameters are used for all models ($\mathcal{F}_{\mathcal{T}}, \mathcal{F}_{\mathcal{S}}, \mathcal{H}_{\mathcal{T}}, \mathcal{H}_{\mathcal{S}}, \mathcal{E}_{\mathcal{S}}, \mathcal{D}_{\mathcal{S}}, \mathcal{E}_{\mathcal{T}}, \mathcal{D}_{\mathcal{T}}$ and $\mathcal{D}_{s\rightarrow t}, \mathcal{D}_{t\rightarrow s}$). In order to represent temporal relations more effectively, we use three consecutive frames as a mini-batch. Our augmentations include a small magnification (from 256 to 272) and random cropping.

A detailed description of our architecture can be found in the Supplementary Materials.

4.3. Baselines

The baselines are state-of-the-art methods for motion retargeting, V2V [44] and EDN [10]. The models are adapted to include an alpha channel as well. EDN is adjusted to work with IUVA (IUV + alpha channel), whereas the V2V model is already optimized for DensePose images, so only one more channel is needed. We train V2V and EDN according to their protocol with the default parameters.

We evaluate the following approaches to assess CTr-GAN's contribution: (a) direct - using trained baselines to map directly from pose to appearance. (b) ours - we use CTrGAN to generate the pose images, then use the baselines to render the appearance.

4.4. Metrics

Gait quality. We evaluate our results by gait recognition models GaitSet [11], GaitGL [28] and GaitPart [16], implemented by the OpenGait [17] package.

We report **target-accuracy** - the percentages of times the gait recognition model identifies the generated gait as the target's gait. Given a set of reference videos $\{\mathbf{I^{s_i}}\}$ and a generated video $\mathbf{I^g}$, the goal is to find the reference video in which the gait pattern is the most similar. For that, references videos are ranked according to their distance from $\mathbf{I^g}$:

$$D_{s_i,g} = \|\mathcal{M}(\mathbf{I^{s_i}}) - \mathcal{M}(\mathbf{I^g})\|_2, \tag{14}$$

where $\mathcal{M}$ is the specific model. The most similar reference video is considered as the one with the highest top-3 (minimum distance) frequency $\mathbf{I^{s_i*}}$. The identified gait is of the subject in $\mathbf{I^{s_i*}}$ and the recovered distance is $D_{s_i*,g}$. In our case the generated video $\mathbf{I^g}$ is $\mathcal{G}_M\left(\mathcal{G}_{s\rightarrow t}\left(\mathbf{P^g}\right)\right)$, where $\mathcal{G}_M$ is the pose to appearance model that is in use (*e.g.* V2V).

We train the models on all the subjects and half of the CASIA-A videos.

Appearance quality. We use the following metrics to evaluate our appearance quality: Inception Score (IS)[34], Structural Similarity (SSIM) [46], Perceptual Image Patch Similarity (LPIPS), [50] and Frechet Inception Distance (FID) [22].

FID and IS metrics measure statistical differences between sets of images and not directly between individual images. However, SSIM and LPIPS evaluate the generated

image based on a single, ground truth image. Due to the fact that the generated synthetic image $\widetilde{\mathbf{I}}_{\mathbf{i}}^{\mathbf{t}}$ in our test set can be derived from an unseen source's pose, a ground-truth image $\mathbf{I}_{\mathbf{i}}^{\mathbf{t}}$ is not always available. We therefore use the Chamfer Distance [6] to recover the nearest ground truth image from the reference video sequence $\{I_k^r\}$:

$$E_{CD} = \frac{1}{N} \sum_{i=1}^{N} \min_k Q(\widetilde{\mathbf{I}}_{\mathbf{i}}^{\mathbf{t}}, \mathbf{I}_{\mathbf{k}}^{\mathbf{r}}), \qquad (15)$$

where Q represents our quality metrics, SSIM or LPIPS. All methods are compared using E_{CD}, both ours and others.

Pose retargeting quality. To estimate the extent to which the basic motion of the source is transferred to the target, the intersection over union (IoU) of the binary silhouettes of the source and target is used. The binary silhouette $\mathcal{S}(\mathbf{P}_{\mathbf{k}}^{\mathbf{A}})$ of subject A in frame k is generated from the alpha channel of its IUVA image I_α with a threshold of $\frac{max(\mathbf{I}_\alpha) - min(\mathbf{I}_\alpha)}{2}$.

Given two sequences A and B with M and N frames respectively, we calculate the average IoU between them according to:

$$R(A, B) = \frac{1}{MN} \sum_{j}^{M} \sum_{k}^{N} IoU(\mathcal{S}(\mathbf{P}_{\mathbf{j}}^{\mathbf{A}}), \mathcal{S}(\mathbf{P}_{\mathbf{k}}^{\mathbf{B}})). \quad (16)$$

We report R where A is the generated target and B is the source, and when A is the generated target and B is its original sequence.

4.5. Experiments

#### 4.5.1 CTrGAN successfully generates the gait of the target.

We train our model on thirteen subjects and use the remaining seven for testing. For each subject, both the training and test sets include two video sequences. In this way, we can include in our evaluation cases where sources were not available during training.

For the test set, we generate video sequences for the trained subjects. We deploy all subjects in the dataset as sources, including those not included in the training set. We test the ability to identify the generated gait as the target's gait by the gait recognition models, before and after applying our approach. Our results demonstrate that our approach can generate a more realistic gait for the target by an order of magnitude than previous methods.

Table 1 presents our main results, the target-accuracy for V2V and EDN. In all the tables, bold represents the best result. The top of the table shows the baselines applied directly to the pose of the source. At the bottom of the table, CTrGAN generates the poses of the target before applying

Method	Model	GaitPart	GaitSet	GaitGL
-	EDN	16.94	29.44	16.67
	V2V	3.89	3.61	4.17
ours	EDN	18.89	**62.78**	36.39
	V2V	**84.72**	56.67	**68.06**

Table 1: The target-accuracy ↑. The top row is before and the bottom row is after applying CTrGAN to generate the poses. Our approach significantly improves the ability to generate the target's gait, up to approximately ×21 than existing methods (GaitPart+V2V).

(a) Before   (b) After
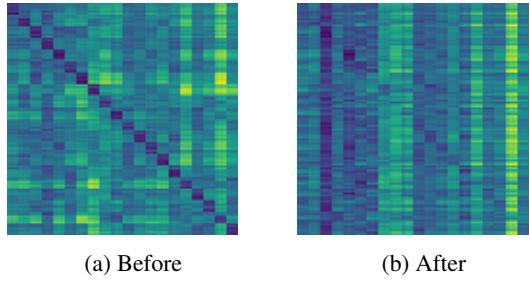
Figure 7: GaitSet's distance matrix for subject three (the target) in the training set before and after applying our method. The darker the color the lower the value. It can be seen that before deploying our model, GaitSet easily distinguishes between the generated and real gait and can identify the true sources. After applying our approach, GaitSet identifies for most of the cases the generated gait as the real gait of subject three.

the baseline methods. It can be seen that for all the methods, CTrGAN significantly improves the ability to generate the natural gait of the target. All models have failed to recognize the gait of the target in the case of V2V without CTrGAN. This implies that indeed, V2V can accurately mimic the movements of the source by the target, in accordance with its original goal. On average, the generated gait rendered by V2V is approximately 17 times more likely to match the target's gait when using our method.

Figure 7 shows the GaitSet's distances for subject number three (the target) with respect to all the twenty subjects in the dataset (the sources). Dark colors represent low values, whereas light colors represent high values. The lower the distance, the more similar the gait in the reference sequence is to that in the generated sequence. Figure 7a presents the distance matrix for the V2V method in the direct approach. It can be seen that GaitSet is able to accurately recognize the sources for all the generated sequences. Figure 7b presents the results for V2V after applying our CTrGAN to generate the poses. For the vast majority of the sequences, GaitSet recognizes the generated gait of the target (subject three) as the real gait.

Model	Features			Target-accuracy ↑
	Attention mechanism	Encoder self-attention	Decoder self-attention	
Cycle Only	✗	✗	✗	5.28
+ Attention	✓	✓	✗	66.21
+ Time-Attention	✓	✓	✓	69.82

Table 2: The target-accuracy of several CTrGAN configurations.

Method	Model	SSIM [CD] ↑	LPIPS [CD]↓	FID↓	IS↓
-	EDN	0.890	0.072	55.79	0.0025
	V2V	0.901	0.063	53.131	0.0010
ours	EDN	0.870	0.101	83.67	0.0030
	V2V	**0.909**	**0.055**	**52.89**	**0.0009**

Table 3: Appearance quality.

Table 3 shows the appearance quality of the different approaches when deploying pose to appearance networks, with and without CTrGAN. Without CTrGAN, the appearance metrics are similar across the different methods. CTrGAN slightly increases the appearance quality for V2V but overall the metrics are comparable. It is expected as the key contribution of CTrGAN is the generation of poses that can naturally be attributed to the target rather than improving the rendering mechanism of an existing pose. A possible explanation for the slight improvement could be that the generated poses by CTrGAN match more naturally the target's appearances that need to be rendered.

The average pose retargeting quality obtained by applying eq. (16) to all the sequences of the generated targets and the sequences of their corresponding sources is 0.8677. The average pose retargeting quality obtained by applying eq. (16) to all the sequences of the generated targets and their original sequences before applying CTrGAN is 0.6325. It is evident that CTrGAN transforms the pose's silhouette of the target to be very similar to the source poses' silhouette.

In the Supplementary Material, we show that existing motion transfer methods retain the gait pattern of the source.

4.5.2 Temporal attention improves accuracy.

Table 2 shows the effect of the different components of CTrGAN on the final results. We employ V2V as our pose-to-appearance network and evaluate the average target-accuracy over all gait recognition models. It can be seen that CycleGAN architecture on its own is not sufficient in order to generate natural poses of the target. Adding encoder self-attention and cross-attention between the image sequence

and the keys using the decoder produces significantly more natural poses. A further improvement is obtained when the decoder self-attention is added, which takes advantage of the temporal relations within the sequence. Additional detailed comparisons can be found in the Supplementary Materials.

5. Detecting Gait Transfer

It is imperative to carefully consider the implications of our method, particularly in light of recent events that are occurring, where misinformation is being used systematically. In introducing the gait transfer problem, we hope to increase awareness of this important issue. We investigate methods for detecting generated gait transfer videos as a first step towards preventing misuse of our approach. In order to identify videos created by our model, we train an appearance-based detector. Choosing an appearance-based detector is practical since the key contribution of our approach is the generation of natural gait patterns whereas the appearance quality is comparable to that of existing motion transfer methods. Our detector classifies videos as either real or generated. The dataset includes both original and generated images of a walking person in an outdoor environment. We use 75 percent of the subjects for training and 25 percent for testing. The deployed model is ResNet152[21]. Using transfer learning, the last FC layer of the model trained on ImageNet is replaced with one that is adapted for two classes, which is fine-tuned on the training data. We achieve an average detection accuracy of 96.2% for held-out target subjects.

6. Conclusion

We introduce a novel approach for gait transfer based on unconstrained videos in-the-wild. We propose quantifiable metrics to better evaluate the quality of the transfer. We present CTrGAN, a novel Transformer-based architecture. Our model cycles between domains by using a series of images and includes self-, cross-, and temporal-attention. We introduce an appearance-based detector and show that it can be highly accurate. Using our approach, we obtained state-of-the-art results.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64, 2020.

[2] Aliakbar Alamdari and Venkat N Krovi. A review of computational musculoskeletal analysis of human lower extremities. *Human Modelling for Bio-Inspired Robotics*, pages 37–73, 2017.

[3] Milene Arantes and Adilson Gonzaga. Human gait recognition using extraction and fusion of global motion features. *Multimedia Tools and Applications*, 55:655–675, 2010.

[4] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

[5] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 122–138. Springer, 2018.

[6] Harry G. Barrow, Jay M. Tenenbaum, Robert C. Bolles, and Helen C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, pages 659–663, 1977.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[10] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8126–8133. AAAI Press, 2019.

[12] Mickael Cormier, Houraalsadat Mortazavi Moshkenan, Franz Lörch, Jürgen Metzler, and Jürgen Beyerer. Do as we do: Multiple person video-to-video transfer. In *4th IEEE International Conference on Multimedia Information Processing and Retrieval, MIPR 2021, Tokyo, Japan, September 8-10, 2021*, pages 84–90. IEEE, 2021.

[13] Adrian Cosma and Ion Emilian Radoi. Wildgait: Learning gait representations from raw surveillance streams. *Sensors*, 21(24):8387, 2021.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[15] Murat Ekinci and Murat Aykut. Human identification using gait. *2006 IEEE 14th Signal Processing and Communications Applications*, pages 1–4, 2006.

[16] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14213–14221. Computer Vision Foundation / IEEE, 2020.

[17] Chao Fan, Chuanfu Shen, and Junhao Liang. Opengait. https://github.com/ShiqiYu/OpenGait, 2022.

[18] Claudio Filipi Gonçalves dos Santos, Diego de Souza Oliveira, Leandro A. Passos, Rafael Gonçalves Pires, Daniel Felipe Silva Santos, Lucas Pascotti Valem, Thierry P. Moreira, Marcos Cleison S. Santana, Mateus Roder, Jo Paulo Papa, and Danilo Colombo. Gait recognition based on deep learning: A survey. *ACM Comput. Surv.*, 55(2), jan 2022.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[20] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7297–7306. Computer Vision Foundation / IEEE Computer Society, 2018.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017.

[23] Drew A. Hudson and Larry Zitnick. Generative adversarial transformers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4487–4499. PMLR, 2021.

[24] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] Muhammad Hassan Khan, Frédéric Li, Muhammad Shahid Farid, and Marcin Grzegorzek. Gait recognition using motion trajectory analysis. In Marek Kurzynski, Michal Wozniak, and Robert Burduk, editors, *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017, Polanica Zdroj, Poland, 22-24 May 2017*, volume 578 of *Advances in Intelligent Systems and Computing*, pages 73–82, 2017.

[26] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[27] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking, 2018.

[28] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14648–14656, October 2021.

[29] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3163–3172, October 2021.

[30] Niki J. Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning (ICML)*, 2018.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H.

Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[32] Jian Ren, Menglei Chai, Sergey Tulyakov, Chen Fang, Xiaohui Shen, and Jianchao Yang. Human motion transfer from poses in the wild. In *European Conference on Computer Vision*, pages 262–279. Springer, 2020.

[33] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos, 2019.

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2377–2386. Computer Vision Foundation / IEEE, 2019.

[36] Luis M Silva and Nick Stergiou. The basics of gait analysis. *Biomechanics and Gait Analysis*, 164:231, 2020.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[38] Jasvinder Pal Singh, Sanjeev Jain, Sakshi Arora, and Uday Pratap Singh. Vision-based gait recognition: A survey. *IEEE Access*, 6:70497–70527, 2018.

[39] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021*, pages 2314–2318. IEEE, 2021.

[40] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[42] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1505–1518, 2003.

[43] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[44] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[46] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[48] Kun Wu, Chengxiang Yin, Zhengping Che, Bo Jiang, Jian Tang, Zheng Guan, and Gangyi Ding. Human pose transfer with disentangled feature consistency. *arXiv preprint arXiv:2107.10984*, 2021.

[49] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based GAN for high-resolution image generation. *CoRR*, abs/2112.10762, 2021.

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018.

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.